7-2009

Exploring inter-concept relationship with context space for semantic video indexing

Xiao-Yong WEI

Yu-Gang JIANG

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Exploring Inter-concept Relationship with Context Space for Semantic Video Indexing

Xiao-Yong Wei
Dept. of Computer Science
City University of Hong Kong
Kowloon, Hong Kong
xiaoyong@cs.cityu.edu.hk

Yu-Gang Jiang
Dept. of Computer Science
City University of Hong Kong
Kowloon, Hong Kong
yjiang@cs.cityu.edu.hk

Chong-Wah Ngo
Dept. of Computer Science
City University of Hong Kong
Kowloon, Hong Kong
cwngo@cs.cityu.edu.hk

ABSTRACT

Semantic concept detectors are often individually and independently developed. Using peripherally related concepts for leveraging the power of joint detection, which is referred to as context-based concept fusion (CBCF), has been one of the focus studies in recent years. This paper proposes the construction of a context space and the exploration of the space for CBCF. Context space considers the global consistency of concept relationship, addresses the problem of missing annotation, and is extensible for cross-domain contextual fusion. The space is linear and can be built by modeling the inter-concept relationship through annotation provided by either manual labeling or machine tagging. With context space, CBCF becomes a problem of concept selection and detector fusion, under which the significance of a concept/detector can be adapted when applied to a target domain different from where the detector is being developed. Experiments on TRECVID datasets of years 2005 to 2008 confirm the usefulness of context space for CBCF. We observe a consistent improvement of 2.8% to 38.8% for concept detection when context space is used, and more importantly, with significant speed-up compared to existing approaches.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Performance, Measurement, Experimentation

Keywords

Context Space, Context-based Concept Fusion, Video Indexing

1. INTRODUCTION

Semantic concept detection (or high-level feature extraction) plays an important role for automatic and interactive video search [12]. Concept detectors (or classifiers) are often developed independently, ignoring the fact that concepts always coexist together and the training examples are naturally multi-labeled. Generally speaking, concepts do not appear in isolation, but are correlated to each other. Such concept correlation provides clues for boosting detection performance, particularly if the robustness of detectors is expected to be uncertain and weak in general. The consensus from multiple contextually related detectors ideally could provide statistical evidence to confirm the existence of a concept. For example, the concept *Car* frequently co-occurs with concepts *Road*. Using the contextual information from *Road* is expected to help detecting *Car*.

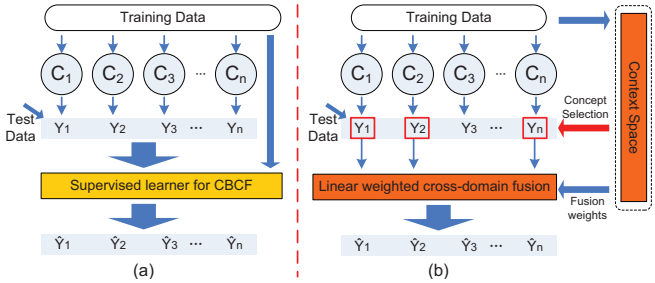


Figure 1: Context-based concept fusion (CBCF) using (a) conventional two-layer learning structure, and (b) the proposed context space.

Improving concept detection performance by utilizing such peripherally related concepts is popularly termed as context-based concept fusion (CBCF), which has received intensive studies in recent years. Most existing works essentially use a two-layer learning structure. As shown in Figure 1(a), n individual concept detectors C_i $(i = 1, \ldots, n)$ are firstly developed. By using detector outputs Y from the first layer as inputs, a supervised learner is then constructed for each concept to get refined prediction $\hat{Y}$. Although performance improvement could be observed as reported in [5][13], there are two major drawbacks. First, these approaches are fully supervised and thus are computationally slow. Second, given a concept, the number of correlated concepts is generally small, compared to the un-correlated ones; using all the concepts as in [13] will significantly affect the performance.

In this paper, we propose a novel approach for CBCF. Instead of relying on the second layer learner, a context space is explicitly built to model the concept relationship. As shown in Figure 1(b), with the context space, CBCF becomes a procedure of concept selection and detector fusion. Specifically, given a target concept, the set of peripherally related detectors are inferred from the context space to jointly boost

CIVR'09 July 8-10, 2009 Santorini, GR

the detection performance. The selected detectors are then linearly fused based on the weights derived from the context space. The weights of detectors can also be adapted for cross-domain fusion, if the peripheral detectors are trained from a domain which is different from the target domain. Context space includes mainly the co-occurrence relationship among concepts. Such information can be learnt from either manual annotations or detectors themselves. For the later case, the learnt context space is capable of adapting concept distribution across different domains and thus facilitating cross-domain contextual fusion. Mathematically, context space is linear, and spanned by a set of basis concepts learnt from concept relationship. Each concept or detector, when projected to this space, becomes a vector facilitating similarity comparison. Given the context space, the similarity between any two concepts is expressed not just by pairwise relatedness, but also with reference to the available basis concepts. In other words, the space is capable of providing a global view of context relationship for inferring a small set of detectors suitable for CBCF. The advantages offered by our proposed approach are summarized as follows:

- *Global versus local measure*: Inter-concept relationship is usually locally determined by pairwise comparison based on observation (e.g., from manual annotations). Such locally determined correlation may not be globally consistent, because the relationship to other concepts is ignored. This paper addresses the problem by building a context space to globally model the concept relationship. Specifically, contextual similarity of two concepts is determined not only based on pairwise comparison, but also considering their relatedness to the basis concepts which form the space. Such uniform measure can greatly facilitate the selection and fusion of concept detectors for CBCF.

- *Incomplete or missing annotation*: Manually labeled concept annotations are always incomplete and forgetful. As reported in [8], missing annotations commonly happen in LSCOM (Large-Scale Concept Ontology for Multimedia) [10], regardless of the efforts being pooled in for labeling of the concepts. A good example is that *Snow* is not labeled together with *Outdoor* in some sample shots by annotators. By pairwise correlation comparison, such missing information can lead to misleading statistics. Using context space for inference, in contrast, the co-occurrence probability of *Snow* and *Outdoor* can still somehow be discovered if *Mountain* is always annotated together with *Outdoor*, and *Snow* is happened to be labeled with *Mountain* in some sample shots. In other words, when concept relationship is modeled as a whole as in context space, the transitivity relationship among concepts can be captured. Missing relationship between two concepts, to certain extent, could still be inferred.

- *Robustness versus randomness*: The fundamental difference between using all available detectors or a subset of detectors for CBCF lies in the level of noise being introduced. For two-layer learner as shown in Figure 1(a), randomness will be introduced if the majority of detectors are not peripherally related to the target concept. Adding the fact that some detectors may not provide accurate estimation, learning a robust detector for target concept which can fully benefit from context inference becomes difficult. Our approach minimizes the randomness from two aspects. First, in view that the number of related detectors is likely to be much less than irrelevant ones, only a very small but helpful set of detectors is picked by querying context space. Second, the significance of a peripheral detector to target concept is known by explicitly inferring from context space. This information provides clues for how to fuse multiple selected detectors, and thus is relatively robust than simply concatenating the scores of detectors as a feature vector for learning of a target detector.

- *Cross-domain contextual fusion*: Context space can be learnt using different sources, either from manual labeling or automatic machine tagging. The later has advantage that no training examples are required. More importantly, using detectors' scores as clues to learn context space allows more realistic fusion of detectors, particularly in the case where the detectors are developed in a domain different from the target domain. We name such fusion strategy as cross-domain fusion. In the experiment, we demonstrate this strategy by online learning of a context space for documentary video domain, based on the detectors trained on news domain. The context space learnt in this way is capable of adapting the distribution of concepts to fit the new domain knowledge, providing a better view of how to fuse concept detectors for CBCF.

- *Speed efficiency*: Knowing the importance (or weight) of a detector towards target concept, our approach adopts linear fusion for CBCF. Apparently, this fusion strategy as illustrated in Figure 1(b) is much more efficient than retraining a classifier as in Figure 1(a).

The idea of building context space was initially proposed in our prior work [15] for video search, in which we named the space as "observability space". In this paper, we further explore context space for CBCF. The remaining sections are organized as follows. Section 2 briefly surveys the existing related work. Section 3 presents the building and modeling of the proposed context space. Section 4 details the utilization of context space for concept fusion in CBCF. Finally, Section 5 shows the experimental results, and Section 6 concludes this paper.

2. RELATED WORK

Semantic concept detection has captured extensive research attention, mainly for its promising role in bridging the semantic gap. The recently released LSCOM (Large-Scale Concept Ontology for Multimedia) [10] includes 834 concepts and a collection of annotations (training examples) for 449 out of the 834 concepts. Based on LSCOM, two detector sets, Columbia374 [17] and VIREO-374 [7][1], are developed and released for public use. Another detector set commonly used is MediaMill-101 [14] which provides 101 concept detectors. These detectors are individually developed, ignoring the inter-concept relationships. Therefore, context-based concept fusion (CBCF) which improves concept detection performance by exploring peripherally related concepts, has

[1]Download site: http://vireo.cs.cityu.edu.hk/research/vireo374/

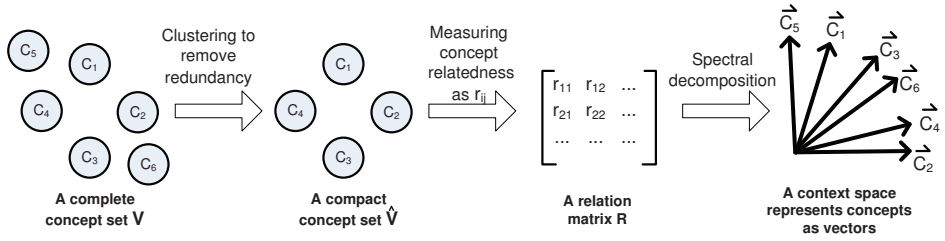


Figure 2: Constructing context space for a given set of concepts by spectral decomposition.

attracted new research attention. CBCF has been investigated in several prior works in recent years [13][5][6][9]. In [13], Smith et al. used SVM to model the contextual relationship. Features for training the SVM is constructed by aggregating the outputs of all the individual detectors. By using this second layer supervised learner, the concept correlation is explored to refine detection results. In [5], Jiang et al. proposed an active CBCF method by firstly soliciting users to annotate a small number of samples. After that, a context-based SVM classifier is learnt which is similar to [13]. In [6], a boosted conditional random field method is proposed for CBCF, in which SVM is used as weak learner for boosting. In [9], Lyndon et al. used inter-concept mutual information based on pseudo-labels to select a set of 75 peripherally related concepts and then trained a SVM for learning the contextual information. These works are all using two-layer learning structure, which combines the individual detector scores of the first layer into feature vectors as inputs for training detectors in the second layer. However, the output of the individual detectors can be unreliable. The detector errors will be propagated to the second layer and therefore degrading the overall performance.

In addition to the two-layer learning techniques, other approaches for CBCF include [11][16]. In [11], Qi et al. proposed a multi-label learning framework derived from Gibbs random field. Though encouraging improvements were observed on TRECVID 2005 data set, the complexity of this method is quadratic to the number of concepts – the computational time to detect 39 concepts is already 25 times longer than that used for training individual classifiers. This prevents its application to a larger set of hundreds of semantic concepts. In [16], Weng et al. proposed a concept fusion method based on graphical model. Through optimizing parameters separately for each concept, an impressive improvement of 16.7% was reported over the VIREO-374 baseline on TRECVID 2006 test set. Nevertheless, all these approaches for CBCF used pairwise comparison of manual annotations to determine inter-concept relationship. This measurement is local as introduced in Section 1. In this paper, we propose to construct a context space which is able to offer global measurement of inter-concept relationship. With the context space, we will show that using a highly efficient linear fusion model with unified parameter setting is able to offer similar or better performance over the existing works.

3. BUILDING CONTEXT SPACE

Given a vocabulary set $\mathcal{V} = [C_1, C_2, \ldots, C_n]$ of n concepts, the aim is to construct a linear space which could effectively model the contextual relationship of concepts in a global way. Figure 2 illustrates the procedure for modeling a context space. Basically concept relationship are modeled and captured in a matrix representation $\mathbf{R}$. The matrix will be further decomposed for deriving the basis vectors which form the linear space. In the first step of this procedure, redundant concepts in the set $\mathcal{V}$ are removed by performing clustering. This eventually results in a compact set of concepts $\hat{\mathcal{V}}$ for producing the matrix $\mathbf{R}$. The inter-concept relationship in $\hat{\mathcal{V}}$ is measured and encapsulated into $\mathbf{R}$. By further performing spectral decomposition on $\mathbf{R}$, a context space which is orthogonal and spanned by basis vectors is constructed. In this space, each concept is represented as a multi-dimensional vector. Concepts not in $\hat{\mathcal{V}}$ can also be projected to the context space by measuring their relatedness with respect to the basis vectors. The similarity of concepts can thus be directly measured by comparing their cosine distance in this space. Note that because the relationship of concepts is globally encoded in the space, concept similarity takes into account not only two concepts under investigation but also their context relatedness to the basis vectors. The details of space construction as shown in Figure 2 will be further described in the remaining subsections.

3.1 Concept Modeling

The matrix $\mathbf{R}$ can be computed from either $\mathcal{V}$ or $\hat{\mathcal{V}}$. Nevertheless, to improve computational stability of spectral decomposition, the compact version $\hat{\mathcal{V}}$ is used instead. $\hat{\mathcal{V}}$ is generated by clustering the n concepts in $\mathcal{V}$. The clustering process will remove redundant concepts in $\mathcal{V}$, resulting in a compact support set $\hat{\mathcal{V}}$ of $m < n$ concepts. $\hat{\mathcal{V}}$ is basically formed by the cluster centroids of $\mathcal{V}$. We adopt agglomerative hierarchical clustering algorithm, and the number of clusters m is determined using inconsistency coefficient measure [4]. Ultimately, the concept relationship in $\hat{\mathcal{V}}$ is encapsulated into the matrix $\mathbf{R} = [\mathbf{r_{ij}}]_{\mathbf{m}\times\mathbf{m}}$.

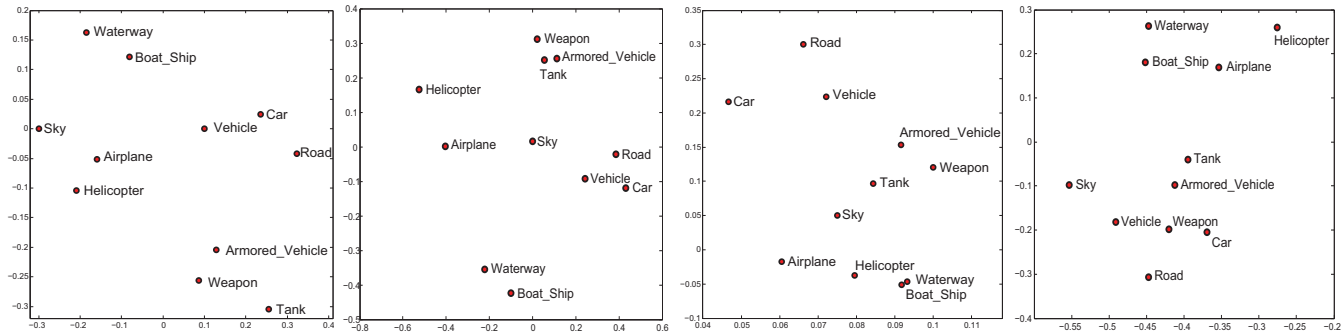
Spectral decomposition [3] is then performed on the matrix $\mathbf{R}$ as following

$$\begin{aligned} \mathbf{R} &= V\Lambda V^T \\ &= (V\Lambda^{\frac{1}{2}}V^T)^T(V\Lambda^{\frac{1}{2}}V^T) \quad (1) \\ &= \mathbf{C}^T\mathbf{C} \quad (2) \end{aligned}$$

where Λ is a matrix with all the eigenvalues of $\mathbf{R}$ on its diagonal, and V is the corresponding eigenvector matrix. The matrix $\mathbf{C}$ encapsulates the concepts in $\hat{\mathcal{V}}$ as vectors. Denote $\mathbf{C} = [\vec{C}_1, \vec{C}_2, \ldots, \vec{C}_m]$, each column of $\mathbf{C}$, $\vec{C}_i$, represents the vector of concept C_i. Theoretically, $\mathbf{C}$ contains the set of basis vectors which span the context space.

3.2 Context Vector Representation

To project an arbitrary concept $u \notin \hat{\mathcal{V}}$ to the concept



(a) (b) (c) (d)

Figure 3: Partial view of concept distribution on LSCOM context space learnt from (a) manual annotations, and (b) detection scores. To contradict, we also show the concept distribution generated by directly using pairwise PM correlation based on (c) manual annotations, (d) detection scores.

space, we perform:

$$\begin{aligned} \mathbf{C}^T \vec{u} &= \mathbf{R}_u \\ \vec{u} &= (\mathbf{C}^T)^{-1} \mathbf{R}_u \end{aligned} \tag{3}$$

where $\mathbf{R}_u$ is a m-dimensional vector, representing the relationship of u to the m concepts in $\vec{\mathcal{V}}$. By the projection, concept u is represented as a vector in the context space, which captures the contextual relationship of u to the m basis vectors. For convenience, we call the vector generated in this way as "context vector".

With vector representation, comparing the similarity of two concepts is computationally efficient. Given the context vectors of concept C_i and C_j, we apply cosine similarity as

$$\lambda_{ij} = Cos(C_i, C_j) = \frac{\vec{C}_i \cdot \vec{C}_j}{|\vec{C}_i||\vec{C}_j|}. \tag{4}$$

where λ_{ij} considers similarity of two concepts with respect to their contextual relatedness to the bases of context space. This measure is relatively robust compared with pairwise measure of concept relatedness.

3.3 Concept Relationship Measurement

When building the context space, we need a measure for quantifying the relationship between two concepts in the matrix $\mathbf{R}$. In this paper, we employ Pearson product-moment correlation (PM) for this purpose. Given two concepts C_i and C_j, the PM correlation is computed as

$$r_{ij} = PM(C_i, C_j) = \frac{\sum_{k=1}^{|\mathcal{T}|}(O_{ik} - \mu_i)(O_{jk} - \mu_j)}{(|\mathcal{T}| - 1)\sigma_i \sigma_j} \tag{5}$$

where O_{ik} indicates the presence of concept C_i in shot k, μ_i and σ_i are the sample mean and standard deviation, respectively, of observing C_i in a data set $\mathcal{T}$. We use two sources of information for computing PM: 1) manual annotations and 2) detection scores. For manual annotations, we set O_{ik} to 1 if C_i is labeled in shot k, and 0 otherwise. When detection scores are used, we simply set O_{ik} equal to the output score (probability) of C_i detector to shot k.

Learning context space using detection scores is flexible, since the responses of detectors can also somewhat reveal the context relationship of concepts. For instance, for shots containing the object "car", the detectors *Car* and *Road* should exhibit consistently higher scores more often than other detectors. Learning such relationship has the advantage that there is no need to annotate all the concepts on the same data set, which is not easy to be achieved in practice.

3.4 Example: LSCOM Context Space

As an example, we use 374 concepts of LSCOM [10] to build a context space and observe the concept distribution. Figures 3(a) and 3(b) show the partial view of context spaces built from manual annotations and detection scores respectively. For illustration purpose, the context space is projected to 2D using MDS (multi-dimension scaling). The concept distributions in 3(a) and 3(b) appear similar to each other, when comparing the relative distance between concepts. The major difference is the concept *Sky*. In 3(a), *Sky* is closer to concepts *Airplane* and *Helicopter*. While in 3(b), *Sky* is more centrally located. This is mainly due to the fact that manual annotations tend to label *Sky* together with *Airplane* and *Helicopter*, but not always for other vehicles. The context space in 3(b) is able to provide a more "objective" view of concept distribution since *Sky* is frequently detected together with vehicles. Using either the manual annotations or detection scores for learning context space has the pros and cons. Detection scores, in contrast to manual annotations, appears as a noisy measure particularly since the robustness of detectors can vary from one to another. For certain groups of concepts, the context space nevertheless is less biased than the one built by using manual annotations. Certainly, the correctness is still dependent on the robustness of detectors.

The context space capable of measuring concept similarity globally, is obtained by representing concepts as vectors through the spectral decomposition on the matrix $\mathbf{R}$ which contains only the local view of pairwise concept relationship. To show the advantages of having global view, we also contrast how concepts distribute by directly measuring the pairwise relationship using PM, as shown in Figures 3(c) and 3(d). There are two main observations when comparing 3(a) and 3(c). First, in 3(c), the distance between *Boat* and *Waterway* appears closer than that between *Car* and *Road*. Such distribution could introduce inconsistency for concept fusion, in a way that the significance of *Waterway* for *Boat* should be similar in par with the importance of *Road* to *Car*. Such consistency is indeed observed in context space where their distances are relatively similar, as shown in 3(a). Context space is capable of keeping such consistency since the basis vectors provide a mean of global reference, in ad-

dition to pairwise correlation, to the two concepts. Second, it is harder to observe the relationship of concepts such as *Airplane*, *Helicopter* and *Boat* to *Vehicle* in 3(c) than in 3(a). We check the manual labels of LSCOM and notice that this is due to the problem of "forgetful annotation": the concept *Vehicle* is more often tagged together with *Car* than with the other three concepts. Context space, nevertheless, is able to capture the relationship of *Vehicle* to other vehicle-related concepts. As shown in 3(a), *Vehicle* is more centrally located and nearby to *Car* as well as *Airplane*, *Helicopter* and *Boat*. We further check the manual labels and notice that context space is able to somehow amend the forgetful annotations due to the fact that concepts such as *Conveyance* and *Transportation_Event* are always labeled together with *Vehicle* and *Airplane*. Context space can pick such hidden cues happened in a transitivity manner and then globally present the concept relationship. Similar observations also happen when comparing Figures 3(b) and 3(d).

4. CONTEXTUAL FUSION

The proposed fusion strategy as shown in Figure 1(b) is composed of two major steps: detector selection and detector fusion. Given a target concept C_t and the learnt context space, the selection procedure starts by first projecting the target concept to context space. Based on the similarities computed by Eq. (4), the top-k concepts with higher similarities to the target concept are then selected as predictors for CBCF. The selected predictors are fused linearly, where the fusion weights are determined by the similarities computed from the context space. Let Y_p be a vector containing the detection scores of an individual detector of concept C_p towards the shots in a test set $\mathcal{T}$, our contextual fusion for a given target concept C_t is conducted as follows

$$\bar{Y}_t = \frac{1}{|\mathcal{P}|} \sum_{C_p \in \mathcal{P}} \lambda_{tp} \cdot Y_p \tag{6}$$

where $\mathcal{P}$ is the set of selected predictor concepts for the target concept C_t, and $\bar{Y}_t$ aggregates the detection scores of all the predictor concepts. The fusion weight λ_{tp} is determined by Eq. (4), which measures the similarity of C_t and C_p in the context space. The final estimation $\hat{Y}_t$ of C_t is obtained by averaging its original detection scores Y_t and the aggregated scores $\bar{Y}_t$

$$\hat{Y}_t = avg(Y_t + \bar{Y}_t). \tag{7}$$

As discussed in Section 3.3, there are two sources of information that can be used to construct the context space, which results in two different λ_{tp}: λ^M_{tp} and λ^S_{tp} respectively derived from the spaces learnt from manual annotations and detection scores. The weight λ^M_{tp} is basically domain specific, while λ^S_{tp} can be adaptive to domain changes and learnt on-the-fly by the time of concept detection. Since the joint appearance of concepts can be different from domain to domain, adaptive learning of a context space for the target domain offers a novel view of detector fusion. Moreover, for the detectors learnt from a domain different from the test data, the context space can amend their significance to other detectors accordingly. For instance, the significance of the concept *Map* to *Weather* can be reduced by adapting the concept distribution in the context space, when the target domain is switched from news to others. Such adaptation greatly facilitates the cross-domain contextual fusion.

Table 1: Performance of CBCF with context space learnt from manual annotations of LSCOM. The VIREO-374 baseline detectors were trained on TV05 development data.

	TV05	TV06	TV07	TV08
VIREO-374 baseline (MAP)	0.303	0.155	0.057	0.040
CBCF (MAP)	0.312	0.179	0.068	0.056
Improvement	2.8%	15.3%	18.2%	38.8%
# of improved concepts	7/10	19/20	16/20	17/19

5. EXPERIMENTS

We conduct experiments using four TRECVID datasets [12]: TV05, TV06, TV07 and TV08, from years 2005 to 2008 respectively. TV05 and TV06 are composed of broadcast news videos in English, Chinese and Arabic. There are 85 hours (45,765 shots) and 150 hours (79,484 shots) of testing videos in TV05 and TV06 respectively. TV07 and TV08 are Dutch videos from the Netherlands Institute for Sound and Vision, containing mainly documentary videos of 50 hours (18,142 shots) and 100 hours (33,726 shots) respectively in the testing sets. For the baseline detector set, we use VIREO-374 [7] which is composed of detectors for 374 LSCOM semantic concepts. The detectors are trained using TRECVID 2005 development set based on the annotations provided by LSCOM. Each detector is associated with three SVM classifiers trained with local interest point features[2], grid-based color moment and wavelet texture respectively. The outputs of the three classifiers are combined as the final detection scores with average fusion.

Following the TRECVID evaluation, we use average precision (AP) to evaluate the results on TV05 and inferred average precision (infAP) for TV06–08. AP approximates the area under precision-recall curve, while infAP estimates the traditional AP when the testing data sets are partially labeled [1]. To aggregate the performance over multiple concepts, we use mean AP for TV05 and mean infAP (MAP) for TV06–08. Throughout our experiments, we report performance on each year's test set over the officially evaluated concepts by NIST. For a given target concept, we uniformly select the three most similar detectors based on the context space for CBCF. Selecting more detectors, especially if the number of detectors is determined adaptively, could possibly lead to better performance. However, we do not aim to elaborate this part in the paper, but rather concentrate on analyzing the effectiveness of context space for CBCF.

We begin by presenting the experimental results for CBCF using the context space learnt from the 374 LSCOM annotations which were manually labeled on TV05 development set (Section 5.1). The results for cross-domain fusion are then described using the context space learnt from the detection scores of VIREO-374 (Section 5.2). Empirical insights about the advantage of the proposed context space (Section 5.3), and performance comparison with existing techniques (Section 5.4) are then presented.

5.1 Context Space for CBCF

Table 1 shows the performance of our CBCF approach with a context space learnt from manual annotations of LSCOM, using VIREO-374 detector set as baseline. Overall, consistent performance improvements (2.8%-38.8%) are

[2]Note that in VIREO-374, only one dictionary of 500 visual keywords was used. Using more dictionaries can lead to better performance as shown in our recent work [2].

Table 2: Selected predictor concepts for the 20 evaluated concepts in TV06 using a context space learnt from manual annotations of LSCOM.

Target Concept	Predictor Concepts
Sports	Athlete, Soccer, Basketball
Weather	Map, Snow, News_Studio
Office	Furniture, Computer, Actor
Meeting	Furniture, Suit, Powerplants
Desert	Weapon, Rocky Ground, Armored Vehicle
Mountain	Hill, Landscape, Sky
Waterscape	Waterway, River, Lake
Corp. Leader	Ties, Face, Single_Person_Male
Police	Police_Security, Police, Military
Military	Military_Personnel, Soldier, Rifle
Animal	Dog, Bird, Horse
Computer-TV-screen	News_Studio, Studio, Studio_Anchor_Person
Flag-US	US Flags, Flags, Speaker_At_Podium
Airplane	Airport, Airplane Flying, Airport_Or_Airfield
Car	Ground Vehicles, Vehicle, Road
Truck	Pickup Truck, Ground Vehicles, Vehicle
People-Marching	Funeral, Crowd, Parade
Explosion Fire	Smoke, Exploding_Ordinance, Weapon
Map	Weather, Studio, News_Studio
Chart	Stock_Market, Sketch, Weather

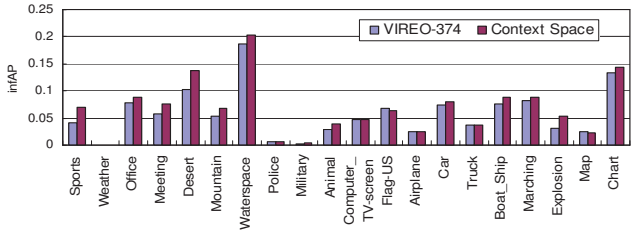


Figure 4: Per-concept performance of CBCF on TV07 test set using context space learnt from manual annotations on TV05 development data (broadcast news videos).

observed on all data sets. More details will be discussed in the following two subsections.

5.1.1 Concept Selection

Selecting appropriate concepts is crucial in our CBCF method. Table 2 lists the top 3 selected predictor concepts for each of the 20 officially evaluated concepts in TV06. We see that the proposed context space is able to help selecting reasonably good concepts for all the target concepts, e.g., {*Athlete*, *Soccer*, *Basketball*} for *Sports*, and {*Hill*, *Landscape*, *Sky*} for *Mountain*. Since the context space is learnt from manual labels on TV05 development data (broadcast news videos), some concepts are selected based on domain specific knowledge. For example, concept *Map* mostly co-occurs with concepts *Weather* and *Studio* (in weather forecast program). These selected concepts are useful for CBCF on news videos, but are useless or even noisy when applied to other data domain such as documentary videos.

To confirm our observations, Figure 4 shows the per-concept performance on TV07 test set using the context space learnt from manual labels on news videos [3]. Compared to VIREO-374 baseline, the performances of 12 out 20 concepts are improved. We see that the fusion performance of concept *Map* is about the same. This is in consistent with our discussions above. The only concept with significant performance drop is *Computer_TV-screen*. This is due to the same reason to concept *Map*, i.e. the selected predictor concepts {*News_Studio*, *Studio*, *Studio_Anchor_Person*} for *Computer_TV-screen* are too domain specific. In Section 5.2, we will show how a context space learnt from detection scores helps in these cases.

[3]Note that 19 out the 20 officially evaluated concepts in TV06 were also used in TV07. The only difference is that *Corp. Leader* in TV06 was replaced by *Boat_Ship* in TV07.

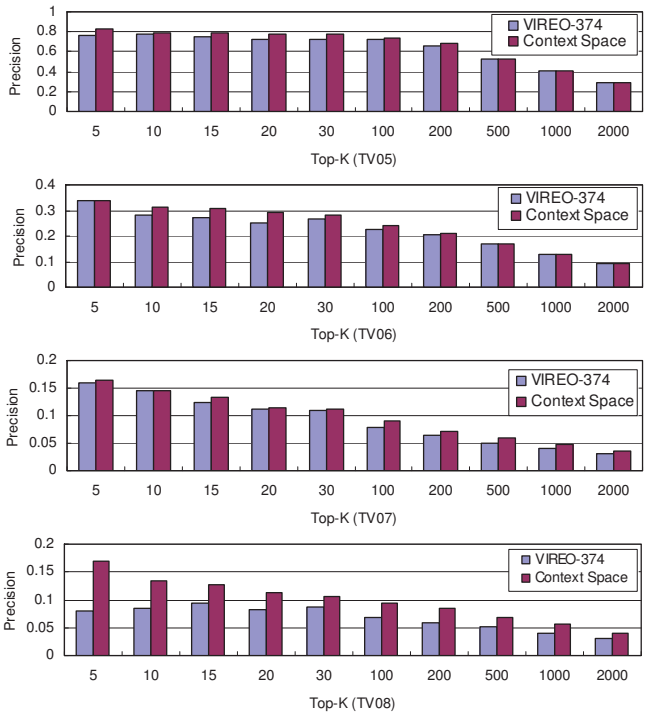


Figure 5: Distribution of true-positive shots (precision@k) over TV05–TV08 test data sets.

5.1.2 Distribution of True-Positive Shots

AP evaluates the performance of a concept detector by approximating the area under precision-recall curve, which essentially reflects the distribution of the true-positive shots in the rank list of the detector. Figure 5 shows precisions at different cutting points of the rank lists for each of the four year's test sets. At each k, we show the mean precision@k over all the evaluated concepts in each year. From the figure, we see that our CBCF method improves the precision at various choices of k for all the test sets. Another interesting observation is that for TV05 and TV06, the CBCF only improves the precision when $k \leq 200$. This indicates that the CBCF only *re-ranks* the top ranked shots and is not able to find more true-positives from the lower part of the list ($k > 2000$). While for TV07 and TV08, we see that the precision@2000 is also significantly boosted. This is probably due to the fact that the VIREO-374 baseline detectors were trained on broadcast news videos (TV05 development set), which are quite different with the documentary videos in TV07 and TV08. This will result in relatively lower baseline performance for TV07 and TV08, and thus leave more room for improvement by CBCF.

5.2 Cross-Domain Contextual Fusion

In this section, we construct context spaces from detection scores for cross-domain contextual fusion (cf. Section 4). Different from Section 5.1, for TV07 and TV08 test sets, we use the classifiers trained on each year's new development data as target detectors. In this scenario, VIREO-374 de-

Table 3: Selected predictor concepts for the 20 evaluated concepts in TV07 using a context space learnt from concept detection scores.

Target Concept	Predictor Concepts
Sports	Walking_Running, Outdoor, Crowd
Weather	Waterscape_Waterfront, Mountain, Hill
Office	Talking, Furniture, Computer
Meeting	Furniture,Powerplants, Suit
Desert	Hill, Mountain, Sky,
Mountain	Mountain, Hill, Landscape,
Waterscape	Boat_Ship, Sky, Waterscape_Waterfront
Boat_Ship	Waterscape_Waterfront, Sky, Waterway
Police	Group, Person, Walking_Running
Military	Military_Personnel, Soldier, People-Marching
Animal	Forest, Dog, Bird
Computer-TV-screen	Office, Comp._Or_Television, Comp._TV
Flag-US	People-Marching, US_Flag, Flag
Airplane	Sky, Airplane_Flying, Helicopter_Hovering
Car	Road, Urban, Ground_Vehicle
Truck	Car, Road, Urban
People-Marching	Crowd, People_Marching, Parade
Explosion_Fire	Explosion_Fire, Exploding_Ordinance, Water
Map	Still_Image, Chart, Sketch
Chart	Still_Image, Charts, Map

Table 4: Performance of cross-domain contextual fusion with context space learnt from detection scores. Note that the baseline detectors are trained on each year's new development data.

	TV06	TV07	TV08
Baseline detectors (MAP)	0.156	0.092	0.119
Cross-domain fusion (MAP)	0.167	0.103	0.123
Improvement	7.3%	11.7%	3.0%
# of improved concepts	16/20	15/20	12/20

tectors will be adopted to enhance target detectors trained in another data domain. A context space is thus learnt for each test set based on the detection scores and then applied for cross-domain contextual fusion.

Table 3 lists the selected predictor concepts for each of the 20 evaluated concepts in TV07 using a context space learnt from detection scores. We can see that this context space is able to pick more reasonable predictor concepts in several cases, e.g., concepts {*Hill*, *Mountain*, *Sky*} are selected for concept *Desert*, while based on the context space learnt from manual labels, {*Weapon*, *Rocky_Ground*, *Armored_Vehicle*} are selected, which is possibly because the highly frequent Iraq war news contain all these concepts.

Table 4 shows the experimental results. Note that although the videos of TV06 are also broadcast news, they are captured in a different year from TV05. Thus we also include TV06 test set in this experiment, in order to compare with the context space learnt from manual annotations. From the Table, we see that the performance improvement for TV06 (7.3%) is not as high as that using a context space learnt from manual annotations (15.3%). This is due to the fact that the context space learnt from detection scores may introduce noises into the concept selection process since many detectors themselves are rather weak and may generate random responds towards testing data. The improvements on TV07 and TV08 are also lower than that on VIREO-374 baseline, which is partially because the new baseline detectors trained on the new data are much stronger. Nevertheless, the overall performance improvement over all the three test sets is still noticeable (3.0%–11.7%). Thus, we conclude that learning context spaces from detection scores is a promising choice since it does not require all the concept being fully annotated, which is hard to be achieved especially when the number of concepts is large or the data domain changes from time to time.

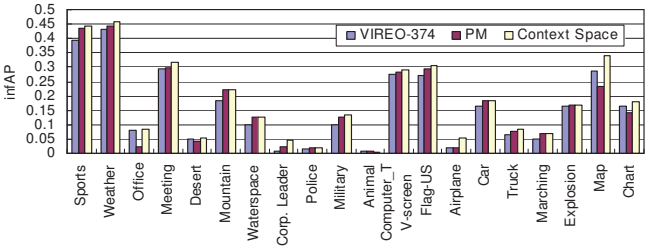


Figure 6: Performance comparison of CBCF using the proposed context space and pairwise PM correlation on TV06. Both context space and PM are computed on manual annotations

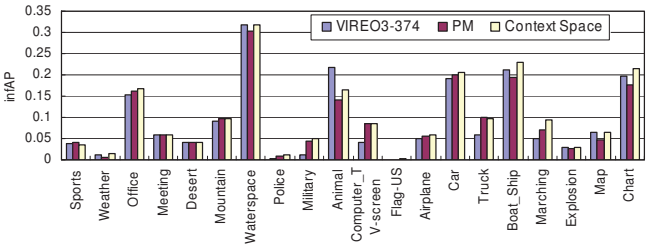


Figure 7: Performance comparison of CBCF using the proposed context space and pairwise PM correlation on TV07. Both context space and PM are computed on detection scores.

5.3 Global vs. Local Measure

In our CBCF method, both concept selection and fusion weights are determined by the context space which is able to offer globally consistent measurement of inter-concept relationships. In this section, we compare the proposed method with pairwise correlation computed by Pearson product-moment (PM; cf. Section 3.3). We also select top-3 predictor concepts for each target concept and use linear weighted fusion for CBCF, where the fusion weights are directly computed by PM.

Figures 6 and 7 show the per-concept performance comparison of CBCF using context space and PM respectively on TV06 test set and TV07 test set. Note that for TV06, both context space and pairwise PM correlation are computed on manual annotations, while for TV07 they are computed on detection scores. From the figures, it is obvious that CBCF with context space is consistently better than that with the pairwise PM correlation, which confirms the advantages of the proposed context space – it is able to offer a global measurement of inter-concept relationship and also can somewhat recover incomplete or missing annotations.

5.4 Performance Comparison and Run Time

In this experiment, we first compare our method to the traditional two-layer learning structure [13][9]. As none of the existing works tested on TV07 and TV08 data sets, in this section we only experiment with TV06 data sets and indirectly compares with other works tested on TV05 data sets. Following [9], for each target concept, we select top-75 peripherally related concepts using the context space and then aggregate the prediction scores of these concepts as a 75-dimensional feature vector for SVM learning. Table 5 re-

Table 5: Performance comparison with two-layer SVM learning on TV06 test set. The best result for each concept is shown in bold.

Concept Name	VIREO-374 Baseline	2-Layer SVM	Context Space
Sports	0.393	0.430	**0.444**
Weather	0.433	0.369	**0.457**
Office	0.081	0.025	**0.082**
Meeting	0.295	**0.328**	0.316
Desert	0.049	**0.059**	0.053
Mountain	0.182	**0.229**	0.220
Waterscape_Waterfront	0.098	**0.146**	0.128
Corporate_Leader	0.008	0.000	**0.047**
Police_Security	0.015	0.015	**0.019**
Military	0.098	0.122	**0.133**
Animal	0.006	0.004	**0.006**
Computer_TV-screen	0.273	0.273	**0.289**
Flag-US	0.269	0.291	**0.306**
Airplane	0.018	0.050	**0.054**
Car	0.164	**0.191**	0.183
Truck	0.064	**0.082**	**0.082**
People-Marching	0.050	0.060	**0.068**
Explosion_Fire	0.165	**0.168**	**0.168**
Map	0.286	0.294	**0.341**
Chart	0.165	**0.205**	0.181
MAP	0.155	0.167	**0.179**
Improvement	–	7.0%	15.3%

ports the results. We see that the proposed CBCF method using context space performs best for 14 out of 20 concepts. Overall, the improvement of the context space based CBCF is 15.3%, which doubles that of the 2-layer SVM learning structure. As reported in [9], their learning structure already outperformed another earlier work [6] on a subset of TV05. Thus we will not directly compare with [6]. The highest performance improvement of the existing CBCF techniques was reported in [16], which improved VIREO-374 baseline by 16.7% on TV06 test set through optimizing a graphical model with fine tuned parameters for each of the 20 concepts. While in our method, we offer similar performance gain by uniformly selecting 3 concepts for all target concepts and using simple but highly efficient linear fusion.

The proposed CBCF method is extremely fast. Learning the context space takes less than 10 seconds on a regular PC using either manual annotation or detection scores (on TV06 test set). Once the context space is built, the CBCF with linear fusion only takes 6 seconds over 20 concepts on TV06 test set containing 79,484 shots. This is much faster than the existing techniques which all involve a computationally expensive optimization process.

6. CONCLUSIONS

We have presented our approach for the construction and exploration of context space for CBCF. The approach turns CBCF into a procedure of concept selection and detector fusion. Under this procedure, our contributions include mainly the proposal of linear space for uniform context measurement, concept selection, and the insights to cross-domain contextual fusion. The building of context space using scores from detectors enlightens the possibility of learning concept relationship without manual labeling, while addressing the issue of cross-domain detection. Such learning is highly efficient and can be conducted on-the-fly as demonstrated in our experiments. The consistent performance improvement being observed when using TRECVID data sets of years 2005 to 2008 also confirm the merit of our approach.

Our current work can be extended in the following ways. In this paper, we only empirically select three most relevant (positive) detectors for CBCF. Using more (or less) detectors, including the negative concepts, remains an issue to be studied. The criteria for selection depends on the factors such as properties of a concept (whether it is general or specific), size of detector set, and also the target domain. In addition, we do observe some fundamental differences between the context spaces learnt from manual annotations and detection scores. Both spaces actually offer different views of concept distributions – human view which could be subjective and forgetful, and machine view which could be noisy. Combining both views for CBCF could be an interesting research study.

7. ACKNOWLEDGMENTS

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 118906), and a grant from City University of Hong Kong (Project No. 7002438).

8. REFERENCES

[1] J. A. Aslam and E. Yilmaz. Inferring document relevance via average precision. In *SIGIR*, 2006.

[2] S.-F. Chang, J. He, Y.-G. Jiang, E. El Khoury, C.-W. Ngo, A. Yanagawa, and E. Zavesky. Columbia University/VIREO-CityU/IRIT TRECVID2008 High-Level Feature Extraction and Interactive Video Search. In *NIST TRECVID Workshop*, 2008.

[3] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[4] A. K. Jain and R. C. Dube. *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[5] W. Jiang, S.-F. Chang, and A. C. Loui. Active context-based concept fusionwith partial user labels. In *ICIP*, 2006.

[6] W. Jiang, S.-F. Chang, and A. C. Loui. Context-based concept fusion with boosted conditional random fields. In *ICASSP*, 2007.

[7] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR*, 2007.

[8] J. R. Kender. A large scale concept ontology for news stories: Empirical methods, analysis, and improvements. In *ICME*, 2007.

[9] L. S. Kennedy and S.-F. Chang. A reranking approach for context-based concept fusion in video indexing and retrieval. In *CIVR*, 2007.

[10] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. G. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86–91, 2006.

[11] G.-J. Qi, X.-S. Hua, Y. Rui, al, and et. Correlative multi-label video annotation. In *ACM MM*, 2007.

[12] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *ACM MIR*, 2006.

[13] J. R. Smith, M. Naphade, and A. P. Natsev. Multimedia semantic indexing using model vectors. In *ICME*, 2003.

[14] C. G. M. Snoek, M. Worring, J. C. Gemert, J.-M. Geusebroek, and Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM MM*, pages 421–430, 2006.

[15] X.-Y. Wei and C.-W. Ngo. Fusing semantics, observability, reliability and diversity of concept detectors for video search. In *ACM MM*, 2008.

[16] M.-F. Weng and Y.-Y. Chuang. Multi-cue fusion for semantic video indexing. In *ACM MM*, 2008.

[17] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia university's baseline detectors for 374 LSCOM semantic visual concepts. Technical report, Columbia University, 2007.